CLT and Confidence Intervals

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's plan

Part 1: Review the asymptotics toolbox from Wednesday

Markov → Chebyshev → LLN → Consistency

Part 2: The Central Limit Theorem

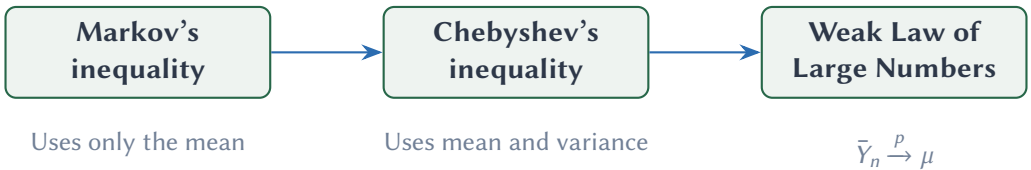
LLN says *where*; CLT says *what shape*

Part 3: Confidence intervals

The payoff: quantifying uncertainty

Wednesday: Midterm.

Review: the proof chain



Each result builds on the last. The entire chain is about four lines of math.

This went fast on Wednesday. Let's walk through it again carefully.

These are bounding tools for reasoning under ignorance

- **Known**: a few summary statistics (mean, maybe variance) of X
- **Unknown**: the full distribution of X
- **Goal**: bound tail probabilities — $\mathbb{P}(X \geq a)$ or $\mathbb{P}(|X - \mu| \geq t)$

Worst-case guarantees that hold for **any** distribution with those moments

No parametric assumptions needed.

This reasoning has a long history

- **Bernoulli** (1713): How many jury trials before you can trust the verdict rate?
- **Insurance**: Upper bound on catastrophic claims, knowing only the average
- **Statistics**: Proving $\bar{X}_n \to \mu$ without knowing the population shape

Today: Markov → Chebyshev → LLN. Each uses one more moment to sharpen the bound.

Two inequalities, two questions

Markov asks:

> *"What fraction of units are above some value a?"*

Note: no reference to the mean. Just: how much of the distribution sits above a?

Chebyshev asks:

> *"What fraction of units are more than k standard deviations from the mean?"*

Now distance from the mean is the question.

Markov's inequality bounds tail probabilities using only the mean

Situation: You run a hospital. You know the average ER bill is \$5,000. You *don't* know the full distribution of bills.

Question: What's the most that could exceed \$25,000?

Markov's Inequality

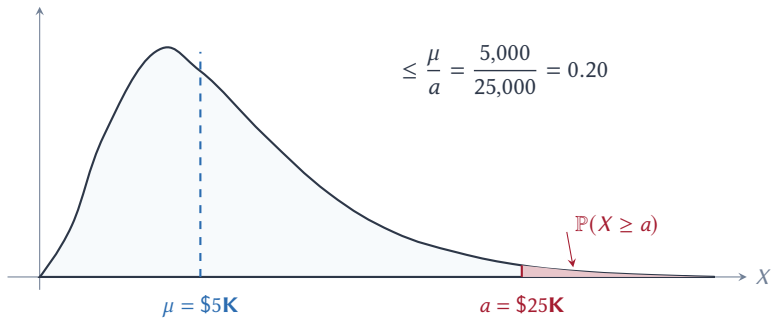
If $X \geq 0$ and $\mathbb{E}[X]$ exists, then for any $a > 0$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Apply: $\mathbb{P}(X \geq \$25{,}000) \leq \frac{5{,}000}{25{,}000} = 0.20$

At most 20% of ER bills exceed \$25K — guaranteed, regardless of the distribution.

Markov bounds the right tail using only the mean



Only uses the mean. Crude — but works for *any* non-negative random variable.

Chebyshev's inequality adds the variance for a tighter bound

Situation: Same hospital. Average ER bill is \$5,000. Now you also know the SD is \$3,000.

Question: What fraction of bills could be *catastrophically* far from the mean — say, more than \$9,000 away?

Chebyshev's Inequality

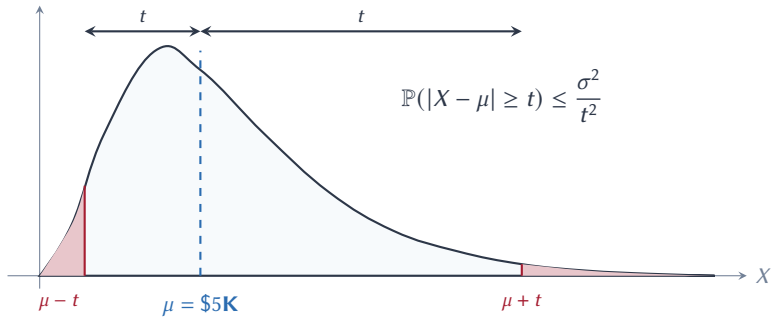
For any random variable X with $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2 < \infty$:

$$\mathbb{P}\big(|X - \mu| \geq t\big) \leq \frac{\sigma^2}{t^2}$$

Let k = distance from the mean in SDs (i.e. $t = k\sigma$):

$$\mathbb{P}\big(|X - \mu| \geq k\sigma\big) \leq \frac{1}{k^2}$$

Chebyshev bounds both tails around the mean



Uses mean *and* variance. Tighter than Markov — same distribution, less area.

Chebyshev example: hospital ER bills

X = ER bill. $\mathbb{E}[X] = \$5{,}000$, $\text{SD}(X) = \$3{,}000$.

Apply: $t = \$9{,}000 = 3\sigma$, so $k = 3$:

$$\mathbb{P}(|X - 5{,}000| \geq 9{,}000) \leq \frac{1}{k^2} = \frac{1}{9} \approx 0.11$$

At most 11% of bills exceed \$14,000 (the lower tail is truncated at zero).

Knowing the variance sharpened the bound — compare Markov's 20% for \$25K to Chebyshev's 11% for \$14K.

Both proofs are short — and Chebyshev comes from Markov

Markov's Inequality

Given: $X \geq 0$, $\mathbb{E}[X]$ exists

Proof:

$$\mathbb{E}[X] = \int_0^\infty x f(x)\, dx$$
$$\geq \int_a^\infty x f(x)\, dx$$
$$\geq a \int_a^\infty f(x)\, dx$$
$$= a\, \mathbb{P}(X \geq a)$$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Chebyshev's Inequality

Given: $\mathbb{E}[X] = \mu$, $\text{Var}(X) = \sigma^2$

Proof:

$$\mathbb{P}(|X - \mu| \geq t)$$
$$= \mathbb{P}\big((X - \mu)^2 \geq t^2\big)$$

Apply Markov to $(X - \mu)^2$:

$$\leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Both hold for ANY random variable X — no distributional assumptions.

Bounding extremes inside a dataset

Setting: Oklahoma tornado counts, 2000–2024.

X = number of tornados in a given year. $\mu \approx 60$

Markov: What fraction of years could have 150+ tornados?

$$\mathbb{P}(X \geq 150) \leq \frac{60}{150} = 0.40$$

At most 40% of years — worst case, guaranteed.

We are reasoning **within** the dataset: bounding how many *rows* can be extreme.

The same logic lets us reason from sample to population

Key insight: Under iid sampling, each *sample* is like a row in a dataset.

- **Within**: rows are observations. Bound extreme *observations*.
- **Across**: rows are samples, each giving a $\bar{Y}_n$. Bound extreme *sample means*.

Chebyshev applied to $\bar{Y}_n$ bounds how far any sample mean can land from μ.

Same inequality, new target — and now we're doing inference.

Chebyshev works *within* a distribution — but also *across* to the population

Within: X = one patient's ER bill

"What fraction of *individual bills* are far from the mean?"

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

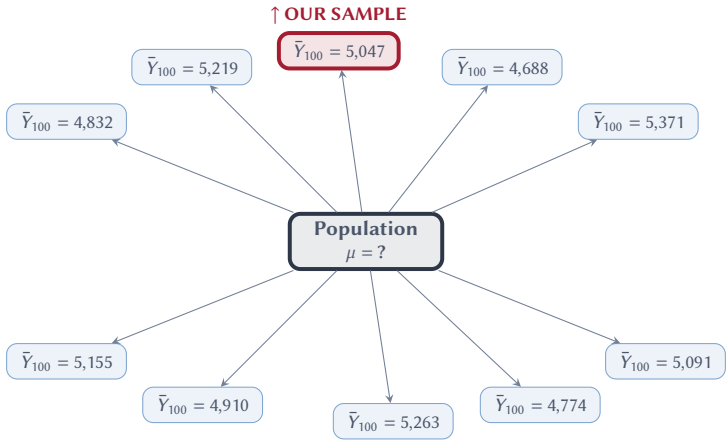
Across: $X = \bar{Y}_n$ (the sample mean is a random variable too!)

"How far can the *sample average* be from the *population* mean?"

$$\mathbb{P}(|\bar{Y}_n - \mu| \geq t) \leq \frac{\sigma^2/n}{t^2}$$

This "across" move — plugging $\bar{Y}_n$ into Chebyshev — is exactly the LLN proof.

Your sample mean is one of many you could have drawn



Every sample of $n = 100$ ER bills gives a different $\bar{Y}_{100}$ — all estimate the same μ.

The power move: plug in $\bar{Y}_n$ for X

Chebyshev works for any random variable — including the sample mean

Chebyshev says: for any X with mean μ and variance σ^2,

$$\mathbb{P}(|X - \mu| \geq t) \leq \sigma^2/t^2$$

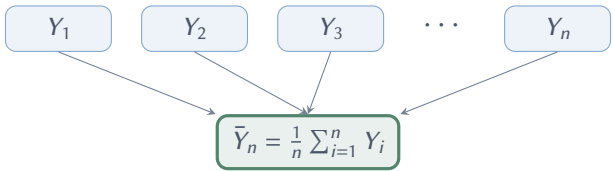
Now let $X = \bar{Y}_n$. Then $\mathbb{E}[\bar{Y}_n] = \mu$ and $\text{Var}(\bar{Y}_n) = \sigma^2/n$.

Substitute:

$$\mathbb{P}(|\bar{Y}_n - \mu| \geq t) \leq \frac{\sigma^2/n}{t^2} = \frac{\sigma^2}{nt^2}$$

As $n \to \infty$, this bound $\to 0$.

A note on subscripts: i and n do different jobs



- Y_i = the ith draw from the population (i = row number)
- $\bar{Y}_n$ = the average of all n draws (n = sample size)

$Y_1, Y_2, \ldots, Y_n$ are the rows. $\bar{Y}_n$ summarizes them into one number.

The Weak Law of Large Numbers follows immediately

WLLN

If $Y_1, Y_2, \ldots$ are i.i.d. with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2 < \infty$, then:

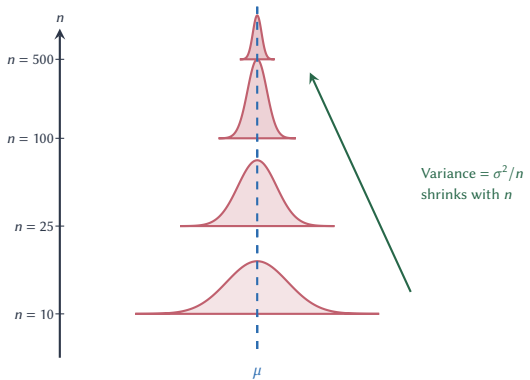
$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$$

Proof: Chebyshev applied to $\bar{Y}_n$:

$$\mathbb{P}(|\bar{Y}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2} \xrightarrow{n\to\infty} 0$$

The sample mean converges in probability to the population mean. This is why statistics works.

As n grows, the distribution of $\bar{Y}_n$ collapses onto μ



At every n, there's a whole distribution of possible $\bar{Y}_n$'s from iid sampling.
As n grows, that distribution squeezes onto μ.

Consistency means the estimator converges to the truth

Definition

$\hat{\theta}_n$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$.

Why does $\bar{Y}_n$ satisfy this?

- Chebyshev bounds how far $\bar{Y}_n$ can be from μ
- The bound is $\sigma^2/(nt^2)$
- $\text{Var}(\bar{Y}_n) = \sigma^2/n \to 0 \quad \Rightarrow \quad \text{bound} \to 0$

That's it. The variance collapses to zero, so the mass concentrates on μ.

Let's pause and see what we just did

1. **Chebyshev**: bounds tail mass for *any* random variable
2. **Within**: applied to individual observations — how many rows can be extreme?
3. **Across**: plugged in $\bar{Y}_n$ for X — same inequality, but now bounding the sample mean's distance from μ
4. **The n**: $\text{Var}(\bar{Y}_n) = \sigma^2/n$ puts n in the denominator of the bound
5. **Consistency**: as $n \to \infty$, the bound $\to 0$, so $\bar{Y}_n \xrightarrow{p} \mu$

Bigger samples → smaller variance → sample mean lands closer to the truth.

Consistency extends to other estimators and functions

By WLLN: $\bar{Y}_n \xrightarrow{p} \mu$

MLE is consistent (under regularity conditions)

CMT: If g is continuous and $\hat{\theta}_n \xrightarrow{p} \theta$, then $g(\hat{\theta}_n) \xrightarrow{p} g(\theta)$

Example: if $\bar{Y}_n \xrightarrow{p} \mu$, then $\log(\bar{Y}_n) \xrightarrow{p} \log(\mu)$

Consistency is the minimum requirement for a useful estimator.

Consistency tells us *where*, but not *what shape*

What LLN gives us:

- $\bar{Y}_n$ converges to μ
- The sampling distribution gets tighter around μ

What LLN does not tell us:

- What *shape* is the sampling distribution?
- How do we build confidence intervals?
- How precise is the estimate for a given n?

For that, we need the CLT.

The population distribution can be anything

X could be distributed:

- Poisson (count of ER visits per day)
- Exponential (time until next tornado)
- Bernoulli (voted or didn't)
- Uniform, skewed, bimodal — anything

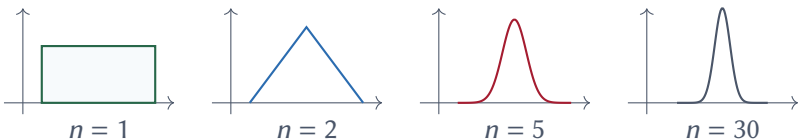
The CLT says: take iid draws, compute $\bar{Y}_n$, repeat.

The distribution of $\bar{Y}_n$ — the *sampling* distribution — will **always become normal**.

No matter what the population looks like.

Averaging makes things normal

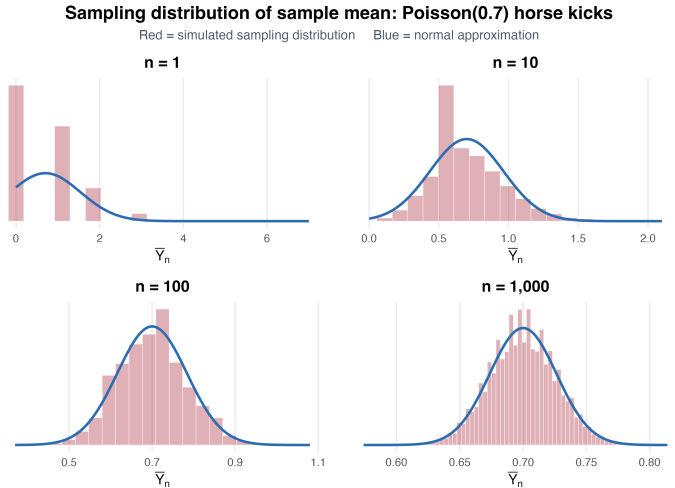
Starting from Uniform[0,1]



As n increases, the distribution of $\bar{Y}$ becomes bell-shaped — no matter where you start.

Even a Poisson population produces a normal sampling distribution

Bortkiewicz (1898): Prussian soldiers killed by horse kicks, $\lambda = 0.7$



What just happened?

Population: Poisson($\lambda = 0.7$) — discrete, right-skewed

- $n = 1$: each "sample" is one corps-year. The distribution is pure Poisson — spikes at 0, 1, 2, 3.
- $n = 10$: average 10 draws. Still skewed, but smoothing out.
- $n = 100$: looks bell-shaped. Normal overlay fits well.
- $n = 1{,}000$: indistinguishable from the normal curve.

The population is Poisson, *but the sampling distribution* is normal. That's the CLT.

So what? Why does this matter for us?

Set aside how remarkable it is that *any* population distribution produces a normal sampling distribution.

Why does this help us as statisticians?

1. **We know the shape.** The normal is fully characterized by μ and σ^2 — so we can write down exact probabilities.
2. **We can build confidence intervals.** If $\bar{Y}_n$ is approximately normal, we know how far it can plausibly be from μ.
3. **We can do hypothesis tests.** We can ask whether an observed $\bar{Y}_n$ is "surprising" under a null.

The LLN told us $\bar{Y}_n$ lands near μ. The CLT tells us *how* near — and gives us the machinery to quantify uncertainty.

Deriving the CLT, step 1: variance of the sample mean

Setup: Draw n observations iid from a population with mean μ and variance σ^2. Repeat many times — each time computing $\bar{Y}_n$.

Question: How much do those sample means vary *across* repeated samples?

Open up $\bar{Y}_n$, pull out the constant, then use independence to sum variances:

$$\text{Var}(\bar{Y}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{n\sigma^2}{n^2}$$

$$\text{Var}(\bar{Y}_n) = \frac{\sigma^2}{n}$$

Deriving the CLT, step 2: standard deviation of the sample mean

From Step 1:

$$\text{Var}(\bar{Y}_n) = \frac{\sigma^2}{n}$$

Take the square root:

$$\text{SD}(\bar{Y}_n) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

This is the $\sigma/\sqrt{n}$ that will appear in the CLT denominator. It's the SD of $\bar{Y}_n$ across samples.

Deriving the CLT, step 3: standardize

To standardize any random variable: subtract its mean, divide by its SD.

$$Z = \frac{\bar{Y}_n - \mathbb{E}[\bar{Y}_n]}{\text{SD}(\bar{Y}_n)}$$

Substitute ($\mathbb{E}[\bar{Y}_n] = \mu$ from iid, $\text{SD}(\bar{Y}_n) = \sigma/\sqrt{n}$ from Step 2):

$$Z = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$$

This Z has mean 0 and variance 1 by construction.

Deriving the CLT, step 4: let n grow

From Step 3:

$$Z = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$$

The CLT says: as $n \to \infty$, this Z converges in distribution to a standard normal.

$$\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

No matter what the population looks like. We saw this with Poisson horse kicks — now here's the math.

The Central Limit Theorem

CLT

Let $Y_1, Y_2, \ldots$ be i.i.d. with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2 < \infty$. Then:

$$\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Equivalent:

$$\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

The "$\xrightarrow{d}$" means the CDF of the left side converges to the standard normal CDF at every continuity point.

The CLT in practice: $\bar{Y}_n$ is approximately normal

For large n:

$$\bar{Y}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Centered at μ (from LLN)
- Standard deviation = $\sigma/\sqrt{n}$ (shrinks with n)
- Shape: **normal** (the new information)

This holds regardless of the original distribution — uniform, exponential, Poisson, whatever.

Only requirements: i.i.d., $\sigma^2 < \infty$.

What does "convergence in distribution" mean?

LLN (convergence in probability): $\bar{Y}_n$ collapses to a *point* — μ.

CLT (convergence in distribution): The standardized Z_n settles into a *shape* — the normal CDF.

Definition

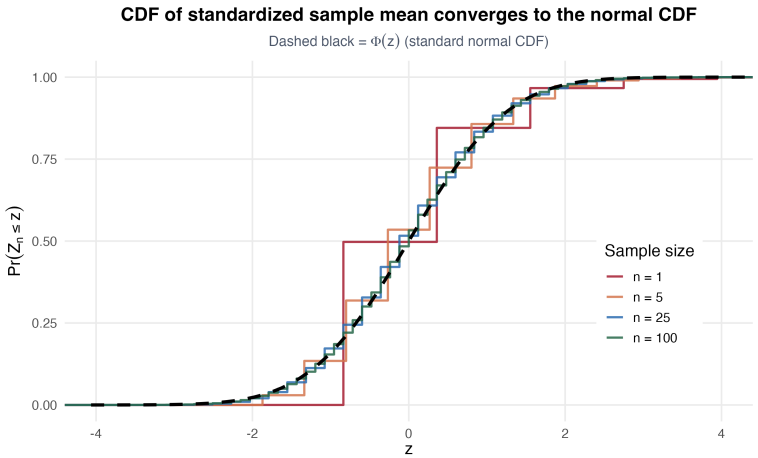
$Z_n \xrightarrow{d} Z$ means: for every value z,

$$\mathbb{P}(Z_n \leq z) \;\rightarrow\; \mathbb{P}(Z \leq z) = \Phi(z)$$

Z_n doesn't converge to a number. It stays random — but its CDF converges to the normal CDF.

You can see it: the CDF converges to the normal

Poisson(0.7) horse kicks, standardized



Two types of convergence, two different results

Convergence in probability ($\xrightarrow{p}$):

- The value itself settles on a number
- $\bar{Y}_n \xrightarrow{p} \mu$: the estimate hits the truth

Convergence in distribution ($\xrightarrow{d}$):

- The *shape* of the distribution stabilizes
- $Z_n \xrightarrow{d} N(0, 1)$: the standardized deviations become normally distributed

$\xrightarrow{p}$ implies $\xrightarrow{d}$, but not vice versa. The CLT uses the weaker form because we're tracking *shape*, not collapse.

Together, $\xrightarrow{p}$ and $\xrightarrow{d}$ unlock inference

$\xrightarrow{p}$ **alone** (LLN): $\bar{Y}_n$ lands near μ. But how near? We can't say.

$\xrightarrow{d}$ **alone** (CLT): The standardized Z_n has a known shape. But shape without convergence isn't useful.

Both together:

1. **Confidence intervals**: 95% of the time, $\bar{Y}_n$ falls within $\pm 1.96 \cdot \sigma/\sqrt{n}$ of μ
2. **Hypothesis tests**: Is the observed $\bar{Y}_n$ surprising under a null?
3. **Sample size planning**: How large must n be to get a CI of a target width?

This is where we're headed next.

Example: presidential approval poll

Population: 0–100 feeling thermometer, $\mu = 45$, $\sigma = 30$, $n = 900$

By CLT:

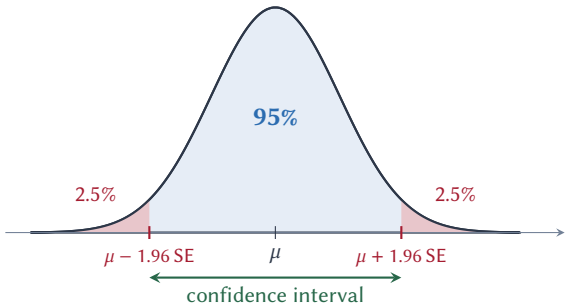
$$\bar{Y}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(45, \frac{30^2}{900}\right) = N(45, 1)$$

Standard error: $\text{SE} = \sigma/\sqrt{n} = 30/\sqrt{900} = 1$

The sampling distribution of $\bar{Y}_n$ is approximately normal, centered on $\mu = 45$, with a SD of 1 point across repeated polls.

The CLT gives us this picture

Sampling distribution of $\bar{Y}_n$ under the CLT



The CLT tells us the shape. 95% of sample means fall within ± 1.96 SEs of μ — that's the confidence interval we're about to build.

How far can the poll be from the truth?

$\bar{Y}_n \approx N(45, 1)$, $\quad \text{SE} = 1$

Question: $\mathbb{P}(|\bar{Y}_n - 45| < 2)$?

Recall: for a normal distribution, $\approx 95\%$ of the mass falls within $\pm$ 1.96 SDs of the mean. Here SE $= 1$, so 2 SDs ≈ 2 points.

$$\mathbb{P}(|\bar{Y}_n - 45| < 2) = \mathbb{P}\left(\left|\frac{\bar{Y}_n - 45}{\text{SE}}\right| < \frac{2}{\text{SE}}\right) \qquad \text{(standardize: divide by SE)}$$

$$= \mathbb{P}\left(\left|\frac{\bar{Y}_n - 45}{1}\right| < \frac{2}{1}\right) \qquad \text{(plug in SE = 1)}$$

$$\approx \mathbb{P}(|Z| < 2) \approx 0.95 \qquad \text{(CLT: this ratio} \approx N(0, 1))$$

95% chance the sample mean is within 2 points of the truth.

Same question, messier numbers

Hospital ER bills: $\mu = \$5{,}000$, $\sigma = \$3{,}000$, $n = 100$, so SE $= 300$

Question: $\mathbb{P}(|\bar{Y}_n - 5{,}000| < 500)$?

Is my sample average within \$500 of the true mean?

$$\mathbb{P}(|\bar{Y}_n - 5{,}000| < 500) = \mathbb{P}\left(\left|\frac{\bar{Y}_n - 5{,}000}{\text{SE}}\right| < \frac{500}{\text{SE}}\right) \quad \text{(standardize)}$$

$$= \mathbb{P}\left(\left|\frac{\bar{Y}_n - 5{,}000}{300}\right| < \frac{500}{300}\right) \quad \text{(plug in SE} = 300\text{)}$$

$$\approx \mathbb{P}(|Z| < 1.67) \approx 0.91 \quad \text{(look up } z = 1.67\text{)}$$

Only 91% — not 95%. \$500 is 1.67 SEs, not 1.96.

Why does averaging produce normality?

Intuition:

- Each Y_i deviates from μ by a random amount
- Positive and negative deviations cancel when averaged
- What remains: tight concentration around μ
- The *shape* of this concentration is always normal

Formal proof: characteristic functions (see A&M §3.2).

How large is "large enough"?

Rules of thumb (heuristics, not guarantees):

Population shape	Typical n needed
Symmetric (uniform, normal)	$n \geq 20$
Moderately skewed	$n \geq 30$
Heavily skewed	$n \geq 50+$
Proportions near 0 or 1	Need $np \geq 5$ and $n(1-p) \geq 5$

A&M simulations show even $n = 100$ can give poor coverage for some distributions. The CLT is not magic — it's an approximation.

CLT for sums (not just averages)

Sometimes we work with sums: $S_n = \sum_{i=1}^{n} Y_i$

- $\mathbb{E}[S_n] = n\mu$
- $\text{Var}(S_n) = n\sigma^2$
- $\text{SD}(S_n) = \sqrt{n}\,\sigma$

CLT for sums:

$$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \xrightarrow{d} N(0, 1)$$

Or: $S_n \approx N(n\mu,\ n\sigma^2)$

Just a rescaled version of the CLT for means. Useful for count data.

Special case: normal approximation to the Binomial

If $X \sim \text{Binomial}(n, p)$, then $X = \sum_{i=1}^{n} Y_i$ where $Y_i \sim \text{Bernoulli}(p)$.

By CLT: $X \approx N(np,\ np(1-p))$ for large n.

Rule of thumb: good if $np \geq 5$ and $n(1-p) \geq 5$.

Example: Flip a fair coin 100 times. $\mathbb{P}(X \geq 60)$?

$X \approx N(50, 25)$:

$$\mathbb{P}(X \geq 60) \approx \mathbb{P}\left(Z \geq \frac{60-50}{5}\right) = \mathbb{P}(Z \geq 2) \approx 0.023$$

The standardization step is always the same

Recipe: To find $\mathbb{P}(\bar{Y} \in \text{some range})$:

1. Write $\bar{Y} \approx N(\mu,\ \sigma^2/n)$
2. Standardize: $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$
3. Look up the standard normal probability

Example: $Y_i \sim \text{Exp}(1/5)$, so $\mu = 5$, $\sigma^2 = 25$. With $n = 100$:

$$\mathbb{P}(\bar{Y} > 5.8) \approx \mathbb{P}\left(Z > \frac{5.8 - 5}{5/\sqrt{100}}\right) = \mathbb{P}(Z > 1.6) = 0.055$$

The CLT requires i.i.d. and finite variance

The CLT fails if:

- **Not i.i.d.**: time series, clustered data, spatial dependence
- **Infinite variance**: heavy-tailed distributions (Cauchy, Pareto with $\alpha \leq 2$)
- n **too small**: approximation hasn't kicked in yet

Extensions exist: CLT variants for dependent data (Lindeberg–Feller), bootstrap for small samples.

The standard error measures the precision of an estimator

By CLT: $\bar{Y}_n \approx N(\mu,\ \sigma^2/n)$

Standard Error

$$\text{SE}(\bar{Y}_n) = \frac{\sigma}{\sqrt{n}}$$

Problem: We don't know σ.

Solution: Estimate it with $\hat{\sigma} = \sqrt{S^2}$ where $S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$

Estimated SE: $\widehat{\text{SE}} = \hat{\sigma}/\sqrt{n}$

But why is plugging in $\hat{\sigma}$ for σ valid?

Slutsky's theorem: plugging in consistent estimators is valid

Slutsky's Theorem

If $T_n \xrightarrow{d} T$ and $S_n \xrightarrow{p} c$, then:

- $T_n + S_n \xrightarrow{d} T + c$
- $T_n \cdot S_n \xrightarrow{d} c \cdot T$
- $T_n / S_n \xrightarrow{d} T/c \quad$ (if $c \neq 0$)

In words: if one sequence has a limiting distribution and the other converges to a constant, they compose nicely.

This is why we can replace unknown parameters with consistent estimates without changing the asymptotic result.

Slutsky in action: replacing σ with $\hat{\sigma}$

By CLT: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$

By LLN: $\hat{\sigma} \xrightarrow{p} \sigma$, so $\hat{\sigma}/\sigma \xrightarrow{p} 1$.

By Slutsky:

$$\frac{\bar{Y}-\mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1)$$

We can use the estimated SE in place of the true SE.

This is what makes confidence intervals possible in practice — we never know σ.

From the CLT to confidence intervals

The payoff of the entire course so far

Since $\frac{\bar{Y}-\mu}{\hat{\sigma}/\sqrt{n}} \dot{\sim} N(0,1)$:

$$\mathbb{P}\left(-1.96 < \frac{\bar{Y}-\mu}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \approx 0.95$$

Rearrange (solve for μ in the middle):

$$\mathbb{P}\left(\bar{Y} - 1.96\,\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{Y} + 1.96\,\frac{\hat{\sigma}}{\sqrt{n}}\right) \approx 0.95$$

This is a 95% confidence interval for μ.

The confidence interval formula

$$95\%\ \text{CI}: \quad \bar{Y} \pm 1.96 \times \widehat{\text{SE}}$$

General formula for level $1 - \alpha$:

$$\bar{Y} \pm z_{\alpha/2} \times \widehat{\text{SE}}$$

Confidence level	α	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

Higher confidence → wider interval. More precision → more data.

The 95% is about the procedure, not any single interval

Correct:

- Repeat the experiment many times, build a CI each time
- 95% of those intervals contain μ

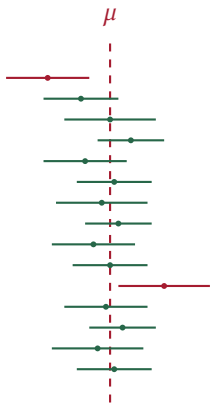
WRONG: "There is a 95% probability that μ is in this interval."

Why wrong:

- μ is fixed — not random
- The *interval* is random (depends on the sample)
- Once computed, it either contains μ or it doesn't

Visualizing the coverage property

Each horizontal line is one CI from a new sample



Green: contains μ. Red: misses μ.

In the long run, 95% of intervals cover the truth.

The rearranging step in detail

From a probability statement about Z to an interval for μ

Start: $\mathbb{P}(-1.96 < Z < 1.96) \approx 0.95$

Substitute $Z = (\bar{Y} - \mu)/(\hat{\sigma}/\sqrt{n})$:

$$\mathbb{P}\left(-1.96 < \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \approx 0.95$$

Multiply through by $\hat{\sigma}/\sqrt{n}$:

$$\mathbb{P}\left(-1.96\,\frac{\hat{\sigma}}{\sqrt{n}} < \bar{Y} - \mu < 1.96\,\frac{\hat{\sigma}}{\sqrt{n}}\right) \approx 0.95$$

Subtract $\bar{Y}$**, multiply by** -1 (flips the inequality):

$$\mathbb{P}\left(\bar{Y} - 1.96\,\widehat{\mathrm{SE}} < \mu < \bar{Y} + 1.96\,\widehat{\mathrm{SE}}\right) \approx 0.95$$

Think of the CI as a net, not a target

What's fixed?	μ (a number, not random)
What's random?	The interval $(\bar{Y} \pm 1.96 \times \widehat{\text{SE}})$
Before sampling:	95% chance the net lands on μ
After sampling:	It either caught μ or it didn't

The "95%" describes the net's quality, not any particular throw.

Example: polling margin of error

Setup: Poll of $n = 1{,}000$ voters. Observed $\hat{p} = 0.52$.

Estimated SE: $\widehat{\text{SE}} = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.52 \times 0.48/1000} = 0.0158$

95% CI: $0.52 \pm 1.96 \times 0.0158 = [0.489,\ 0.551]$

Margin of error: $1.96 \times 0.0158 \approx 0.031$ (± 3.1 percentage points)

This is the "$\pm 3\%$" you see in news reports of polls.

What determines the width of a confidence interval?

$$\text{Width} = 2 \times z_{\alpha/2} \times \frac{\hat{\sigma}}{\sqrt{n}}$$

Three levers:

- **Confidence level** ($z_{\alpha/2}$): Higher confidence → wider interval
- **Variability** ($\hat{\sigma}$): More noise in the data → wider interval
- **Sample size** (n): More data → narrower interval (by $1/\sqrt{n}$)

To halve the width, you need 4× the data. Precision is expensive.

Worked example: CI for a treatment effect

GOTV experiment: $n_T = 500$ treatment, $n_C = 500$ control.

$\bar{Y}_T = 0.65$, $\bar{Y}_C = 0.60$, so $\hat{\tau} = 0.05$

Step 1: Estimated SE (assuming independence):

$$\widehat{\text{SE}}(\hat{\tau}) = \sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}} = \sqrt{\frac{0.2275}{500} + \frac{0.24}{500}} = 0.0305$$

Step 2: 95% CI: $0.05 \pm 1.96 \times 0.0305 = [-0.010,\ 0.110]$

The CI includes zero — we cannot rule out no effect at the 5% level.

Three common mistakes about confidence intervals

Mistake	Correct
"95% probability $\mu \in [a, b]$"	95% of intervals contain μ
"Overlapping CIs $\Rightarrow$ no difference"	CIs can overlap when the difference is significant
"Wider CI = worse estimate"	Width reflects n and σ, not quality

For comparing groups, always use the CI for the *difference*.

Sample size planning: how much data do you need?

Goal: 95% CI with margin of error $\leq m$

Requirement: $1.96 \times \sigma/\sqrt{n} \leq m$

Solve:

$$n \geq \left(\frac{1.96\,\sigma}{m}\right)^2$$

Example: Poll with $\pm 3\%$ margin of error, $\sigma \approx 0.5$ (worst case for proportions):

$$n \geq \left(\frac{1.96 \times 0.5}{0.03}\right)^2 = 1{,}067$$

This is why national polls use $n \approx 1{,}000$.

The Delta Method: CLT for transformations

What if we care about $g(\mu)$, not μ itself?

Delta Method

If $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$ and g is differentiable at μ with $g'(\mu) \neq 0$:

$$\sqrt{n}\big(g(\bar{Y}) - g(\mu)\big) \xrightarrow{d} N\big(0,\ [g'(\mu)]^2\sigma^2\big)$$

In practice:

$$g(\bar{Y}) \approx N\left(g(\mu),\ [g'(\mu)]^2\frac{\sigma^2}{n}\right)$$

Smooth transformations of asymptotically normal estimators are also asymptotically normal.

Delta method example: log odds

Setup: Estimate the log odds $\theta = \log\left(\frac{p}{1-p}\right)$ from binary data.

By CLT: $\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0,\ p(1-p))$

Let $g(p) = \log(p/(1-p))$. Then $g'(p) = \frac{1}{p(1-p)}$.

By Delta Method:

$$\sqrt{n}\big(g(\hat{p}) - g(p)\big) \xrightarrow{d} N\left(0,\ \frac{1}{p(1-p)}\right)$$

The SE for log odds is $1/\sqrt{n \cdot \hat{p}(1-\hat{p})}$ — the same formula you'll see in logistic regression output.

The MLE is asymptotically normal

The big connection: Fisher information + CLT

Asymptotic Normality of MLE

Under regularity conditions:

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

Equivalently:

$$\hat{\theta}_{\text{MLE}} \approx N\left(\theta_0, \frac{1}{n\, I(\theta_0)}\right)$$

The variance of the MLE achieves the Cramér–Rao lower bound — the MLE is **efficient**.

This connects Fisher information from 06a to the CLT from today.

Why is the MLE asymptotically normal?

Sketch of the argument

The score function: $s(\theta) = \ell'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i \mid \theta)$

Key facts from 06a:

- $\mathbb{E}[s(\theta_0)] = 0$ (score has mean zero at truth)
- $\text{Var}(s(\theta_0)) = n\, I(\theta_0)$ (Fisher information)

By CLT: $s(\theta_0)/\sqrt{n} \xrightarrow{d} N(0,\ I(\theta_0))$

Taylor expand the FOC $s(\hat{\theta}) = 0$ around θ_0:

$$0 \approx s(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)$$

Solve: $\hat{\theta} - \theta_0 \approx -s(\theta_0)/\ell''(\theta_0)$

The CLT on s and LLN on ℓ'' give the result.

Confidence intervals from the MLE

From asymptotic normality:

$$\hat{\theta}_{\text{MLE}} \approx N\left(\theta_0, \frac{1}{n\, I(\theta_0)}\right)$$

Estimated SE: $\widehat{\text{SE}} = 1/\sqrt{n\, I(\hat{\theta})}$

95% CI:

$$\hat{\theta}_{\text{MLE}} \pm 1.96 \times \frac{1}{\sqrt{n\, I(\hat{\theta})}}$$

This is how statistical software computes confidence intervals for regression coefficients, logit models, and virtually every parametric estimator.

Example: CI for the Poisson rate

Continuing the protest-count example from 06b

$X_i \sim \text{Poisson}(\lambda)$: protests per country-month

From 06b: $\hat{\lambda}_{\text{MLE}} = \bar{X}, \quad I(\lambda) = 1/\lambda$

Asymptotic distribution: $\hat{\lambda} \approx N(\lambda,\ \lambda/n)$

Estimated SE: $\widehat{\text{SE}} = \sqrt{\hat{\lambda}/n} = \sqrt{\bar{X}/n}$

Data: $n = 200$ country-months, $\bar{X} = 3.4$ protests

$$\widehat{\text{SE}} = \sqrt{3.4/200} = 0.130$$

95% CI: $3.4 \pm 1.96 \times 0.130 = [3.15,\ 3.65]$

Example: CI for the Normal mean and variance

Continuing the vote-share example from 06b

$X_i \sim N(\mu, \sigma^2)$: district-level vote shares

From 06b: $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$

CI for μ: Use $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ (unbiased)

$$\bar{X} \pm 1.96 \times \frac{S}{\sqrt{n}}$$

Data: $n = 435$ districts, $\bar{X} = 0.53$, $S = 0.15$

$$\widehat{\mathsf{SE}} = 0.15/\sqrt{435} = 0.0072$$

95% CI: $0.53 \pm 1.96 \times 0.0072 = [0.516,\ 0.544]$

Example: CI for voter turnout via Fisher information

Continuing the voter turnout example from 06a

$X_i \sim \text{Bernoulli}(\theta)$: voter turnout.

From 06a: $\hat{\theta}_{\text{MLE}} = \bar{X} = 0.34, \quad I(\theta) = \frac{1}{\theta(1-\theta)}$

Estimated Fisher information: $I(\hat{\theta}) = \frac{1}{0.34 \times 0.66} = 4.456$

Estimated SE: $\widehat{\text{SE}} = \frac{1}{\sqrt{n \cdot I(\hat{\theta})}} = \frac{1}{\sqrt{200 \times 4.456}} = 0.0335$

95% CI: $0.34 \pm 1.96 \times 0.0335 = [0.274,\ 0.406]$

Identical to $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$ — the Fisher info formula gives the same SE.

Summary: CIs for the distributions we've studied

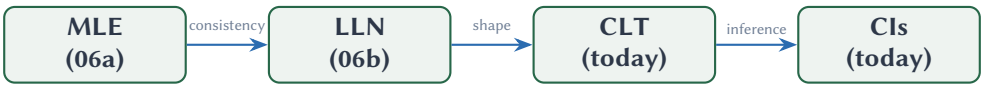
Model	$\hat{\theta}$	$I(\theta)$	$\widehat{\text{SE}}$
Bernoulli(θ)	$\bar{X}$	$\frac{1}{\theta(1-\theta)}$	$\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
Poisson(λ)	$\bar{X}$	$\frac{1}{\lambda}$	$\sqrt{\frac{\bar{X}}{n}}$
Normal(μ, σ^2)	$\bar{X}$	$\frac{1}{\sigma^2}$	$\frac{S}{\sqrt{n}}$

95% CI: $\hat{\theta} \pm 1.96 \times \widehat{\text{SE}}$ in every case.

Key takeaways

1. **LLN**: $\bar{Y}_n \xrightarrow{p} \mu$ (the estimate converges to the truth)
2. **CLT**: $\bar{Y}_n \approx N(\mu,\ \sigma^2/n)$ (the shape is normal)
3. **Slutsky**: replacing σ with $\hat{\sigma}$ is valid asymptotically
4. **CI**: $\bar{Y} \pm z_{\alpha/2} \times \widehat{\text{SE}}$ (quantifying uncertainty)
5. **MLE**: asymptotically normal, efficient, SE from Fisher information
6. **Delta method**: smooth transformations preserve asymptotic normality

The full picture: from data to inference



- **06a**: Build estimators (MLE), measure their quality (bias, variance, Fisher info)
- **06b**: Show they converge to the truth (Markov → Chebyshev → LLN)
- **Today**: Determine the shape (CLT) and quantify uncertainty (CIs)

Wednesday: Midterm

Coverage: Weeks 1–6 (through Estimation)

Format: In-class, closed book

What to know:

- Probability rules, conditional probability, Bayes' theorem, LOTP
- Random variables, PMFs, CDFs, independence
- $\mathbb{E}[X]$, $\text{Var}(X)$, SD, LOTUS, Jensen's inequality
- Named distributions: Bernoulli, Binomial, Poisson, Uniform, Normal, Exponential
- Joint distributions, covariance, correlation, CEF, LIE, LOTV
- Estimation: estimand, estimator, estimate, plug-in, bias, variance, MSE

Today's material (asymptotics, CLT, CIs) is **not** on the midterm.

Good luck on Wednesday

Office hours: Today after class, 3:00–5:00 PM

Kaixiao's section: Friday review session

You have all the tools.

Probability → distributions → expectation → estimation

Each week built on the last.